Development and Initial Validation of the Session Reactions Scale–5 (SRS-5): A Brief Client-Report Measure of In-Session Impact for Feedback-Informed Treatment

Ben Buchanan[1,2], David Hegarty[1,3], and Tomáš Řiháček[4]

[1] *NovoPsych Psychometrics, Melbourne, Victoria, Australia*
[2] *Monash University, Melbourne, Victoria, Australia*
[3] *Southern Cross University, Coffs Harbour, New South Wales, Australia*
[4] *Department of Psychology, Faculty of Social Studies, Masaryk University, Brno, Czech Republic*

Abstract

Objective. Feedback-informed treatment (FIT) often involves a brief, repeatable client-report measure that can be administered after every session. Existing session measures emphasise the therapeutic alliance but they suffer from ceiling effects; fewer index the client's experienced in-session impact. This study reports the development and initial validation of the Session Reactions Scale–5 (SRS-5), a five-item client-report measure of helpful in-session reactions derived from the Session Reactions Scale–3–Brief (SRS-3-B). **Method.** Data were drawn from NovoPsych, the de-identified research dataset of the NovoPsych digital assessment platform, comprising 975 SRS-3-B administrations from 975 clients seen by 50 clinicians, predominantly in Australia. All clinicians consented to the anonymous use of their data for research. Five items were retained to represent the principal domains of helpful session reaction, excluding items that overlapped with the therapeutic alliance, showed marked ceiling effects, or applied only to some sessions. Internal consistency, score distribution, percentile norms, and convergent validity against the Agnew Relationship Measure–5 (ARM-5) were examined. **Results.** The SRS-5 (range 5–25) showed strong internal consistency (α = .86, ω = .86), an approximately normal, mildly negatively skewed distribution (M = 16.9, SD = 4.4, Mdn = 17, skew = −.32), and minimal floor (0.6%) and ceiling (3.7%) effects. In a matched subsample (n = 68 clients; 39 same-day), the SRS-5 correlated r = .41 (95% CI [.18, .59]) with ARM-5 total alliance, rising to r = .63 among same-day pairs. **Conclusions.** The SRS-5 is a reliable, well-distributed ultra-brief measure whose scores converge with, but are not redundant with, the working alliance, supporting its use as a feasible session-by-session index of helpful client impact in routine care.

Keywords: feedback-informed treatment; routine outcome monitoring; session reactions; therapeutic alliance; psychometrics; measurement-based care

Introduction

Feedback-informed treatment (FIT) and the broader practice of routine outcome monitoring refer to the systematic collection of client-report data during the course of psychotherapy and the use of that information to guide ongoing clinical decisions. A growing body of evidence indicates that giving clinicians structured, real-time client feedback improves treatment outcomes, reduces deterioration, and lowers drop-out, particularly for clients who are not progressing as expected (de Jong et al., 2021; Lambert & Shimokawa, 2011). Two classes of feedback are typically distinguished: outcome feedback, which tracks symptom change between sessions, and process feedback, which captures the client's experience of the session itself.

Process feedback in FIT has been dominated by measures of the therapeutic alliance. The alliance, conceived by Bordin (1979) as the bond between client and therapist together with their agreement on the goals and tasks of therapy, is among the most robust predictors of outcome across modalities (Flückiger et al., 2018; Norcross & Lambert, 2018). The most widely used session-level alliance measure in FIT is the Session Rating Scale (SRS; Duncan et al., 2003; Miller et al., 2006), a four-item visual-analogue tool rating relationship, goals and topics, approach or method, and an overall judgement of the session. The Agnew Relationship Measure (ARM; Agnew-Davies et al., 1998) and its short forms (Cahill et al., 2012) provide a complementary, psychometrically grounded account of alliance structure. Many alliance measures, however, suffer from ceiling effects where the scales are not able to discriminate effectively at the top end. This could be due to the fact that it is confronting for clients to respond honestly about their reactions to the therapist themselves.

Alliance, however, is not the only clinically meaningful product of a session. Drawing on Elliott's (1985) empirical taxonomy of helpful and hindering in-session events, Řiháček et al. (2024) developed the Session Reactions Scale–3 (SRS-3) to measure the specific reactions a client has to a session, such as gaining new

insight, feeling more empowered or hopeful, clarifying goals, acquiring coping strategies, and experiencing emotional relief, alongside hindering reactions such as feeling stuck or pressured. The brief version (SRS-3-B) comprises a 10-item Helpful Reactions subscale and a 4-item Hindering Reactions subscale. These impact-focused reactions are conceptually distinct from the alliance: a session may be experienced as relationally warm yet low in impact, or vice versa.

For session-by-session use in busy routine settings, even a 14-item instrument can impose meaningful response burden. There is therefore a case for an ultra-brief, positively keyed measure that isolates the core helpful in-session impact reactions, retains strong measurement properties, and can be completed in well under a minute. The present study develops such a measure, the Session Reactions Scale–5 (SRS-5), from the Helpful Reactions items of the SRS-3-B, and reports its score distribution, internal consistency, normative percentiles, and convergence with the ARM-5.

Method

Data Source and Participants

The de-identified research dataset was from NovoPsych digital psychometric platform. All clinicians in the dataset consented for their anonymised assessment data to be used for research purposes, and no identifying information was accessed for this study. The dataset comprised 975 administrations of the SRS-3-B completed between October 2023 and July 2026 by 975 distinct clients under the care of 50 clinicians (84.0% psychologists, 8.0% counsellors, and 8.0% social workers). Each client contributed only their first administration, so observations were independent. In total, these clients had completed a median of one SRS-3-B administration ($M = 2.7$, range 1–36).

The sample was predominantly Australian: 73.3% of administrations were recorded by Australia-based clinicians, with the remainder from the United States (11.5%), New Zealand (8.1%), the Philippines (4.7%), and other locations (2.4%). Clients were 50.8% female, 27.6% male, and 3.7% gender diverse or other, with 17.9% not specifying gender. Age was recorded for 89.1% of clients, and the median age band was 25–34 years. The SRS-3 is intended for adults aged 18 and over; 9.2% of clients with a recorded age were younger than this. Because the data arise from routine clinical practice rather than a recruited research cohort, the sample is best characterised as a large, naturalistic, predominantly Australian help-seeking population.

Measures

Session Reactions Scale–3–Brief (SRS-3-B). The SRS-3-B (Řiháček et al., 2024) is a 14-item client-report post-session measure with two subscales, Helpful Reactions (10 items) and Hindering Reactions (4 items). Each item is rated from 1 (not at all) to 5 (very much) according to how well the description fits the client's experience of the most recent session.

Agnew Relationship Measure–5 (ARM-5). The ARM-5 (Cahill et al., 2012) is a five-item self-report measure of the therapeutic alliance yielding a total alliance score and bond, partnership, and confidence components. Items are rated on a seven-point scale; one item is reverse keyed. The total alliance mean served as the primary alliance criterion.

Item Selection and the SRS-5

The SRS-5 was formed from five Helpful Reactions items chosen to represent distinct domains of helpful in-session impact identified in the session-impact literature (Elliott, 1985; Řiháček et al., 2024): newly gained insight (Item 1), increased empowerment or hope (Item 4), clarification of problems or goals (Item 7), acquisition of skills or coping strategies (Item 9), and emotional relief (Item 13). Items were retained for content breadth across these domains together with favourable classical item properties (acceptable item–total discrimination and wide response dispersion without strong ceiling effects). The SRS-5 total is the sum of the five items (range 5–25); a mean item score (range 1–5) can also be reported.

Items were selected for the measure's intended use as well as for their statistics. Because the SRS-5 is meant to be completed after every session, often alongside an alliance measure, retained items had to (a) describe the client's experienced in-session impact rather than an appraisal of the therapist, (b) be relevant to most sessions

regardless of orientation or session focus, and (c) discriminate between sessions without marked ceiling effects. Within these criteria, items were chosen to cover as many distinct types of helpful impact as possible. Properties of all ten Helpful Reactions items are given in Appendix Table A1.

Items 2 ("I feel understood, supported, or reassured by my therapist") and 11 ("I feel personally invested in what I need to do in therapy to achieve my goals") were excluded because their content overlaps with the alliance; in Elliott's (1985) terms they are interpersonal rather than task impacts. Both were among the items most strongly related to global alliance in the original validation (Řiháček et al., 2024), and in the matched ARM-5 subsample Item 2 had the strongest correlation with ARM-5 total alliance (r = .49, vs. .14 to .40 for the retained items; Table 1). They also had by far the largest ceiling effects (65.4% and 41.7% rated "Very much", vs. 14.9% to 21.7% for the retained items). Item 7 was retained despite its link to alliance in the original validation because it describes the client's own clarity about problems or goals, and its ARM-5 correlation (r = .38) fell within the range of the other retained items.

Items 3 (feeling more distanced from feelings, thoughts, or memories) and 6 (feeling more hopeful about another person) were excluded because they apply only to some sessions, such as those using defusion or focusing on relationships, and had the highest rates of "Not at all" responses (21.6% and 15.7%, vs. 2.6% to 9.9% for the retained items). Item 3 was also the weakest indicator of the Helpful factor (loading = .38), and Item 6 overlapped with Item 4 (r = .56). Item 14 (acceptance) performed well but overlapped with Items 4 and 13 (r = .55 and .60), so Item 1 (new insight) was kept instead to represent a cognitive type of impact. This cost little: the five highest-loading items would have given α = .87 rather than .86, but more maximum scores (4.4% vs. 3.7%).

The Hindering items were not considered. They form a separate factor (Řiháček et al., 2024) and showed severe floor effects (70.2% to 88.9% rated "Not at all"), so a single, positively keyed total was preferred. Clinicians who want feedback on hindering reactions or on specific reaction types can continue to use the full SRS-3-B.

Analytic Strategy

Internal consistency was estimated with Cronbach's alpha (Cronbach, 1951), McDonald's omega (ω, from a single-factor model; McDonald, 1999), and corrected item–total correlations. Omega is reported alongside alpha because alpha assumes that every item loads equally on the construct, which is rarely the case, and it can misestimate reliability when that assumption fails (Dunn et al., 2014). The score distribution was summarised with descriptive statistics, skew, and floor/ceiling percentages, and percentile ranks were tabulated across the full range. Convergent validity was tested by linking SRS-5 and ARM-5 records on a de-identified client identifier; because both are session-level measures, each client contributed a single pair, formed from the SRS-3-B administration and the ARM-5 administration closest in time to it, yielding independent observations. Pearson and Spearman coefficients (with 95% confidence intervals [CIs] via Fisher's z) were computed for the full matched sample and for the subset completed on the same day. Correlation magnitudes were interpreted following Cohen (1988).

All analyses were conducted in R (version 4.2.0; R Core Team, 2022). Cronbach's alpha, McDonald's omega, corrected item–total correlations, descriptive statistics, and the single-factor principal-axis factor analysis, including regression-based factor scores, were computed with the psych package (version 2.3.9; Revelle, 2023). The intraclass correlation coefficient (ICC) was computed with the irr package (version 0.84.1; Gamer et al., 2019). Pearson and Spearman correlations, with Fisher's z confidence intervals, were computed with the base R function cor.test. Figure 1 was produced with ggplot2 (version 3.4.4; Wickham, 2016).

Results

Internal Consistency and Item Properties

The five SRS-5 items showed strong internal consistency (Cronbach's α = .86; McDonald's ω = .86). Corrected item–total correlations ranged from .59 (Item 1) to .72 (Item 9), all within the acceptable-to-strong range, indicating that each item contributed coherently to the composite (Table 1). Item 1 (new insight) was the weakest contributor, consistent with its broader, less impact-specific content.

Table 1
SRS-5 Item Content and Classical Item Properties (N = 975)

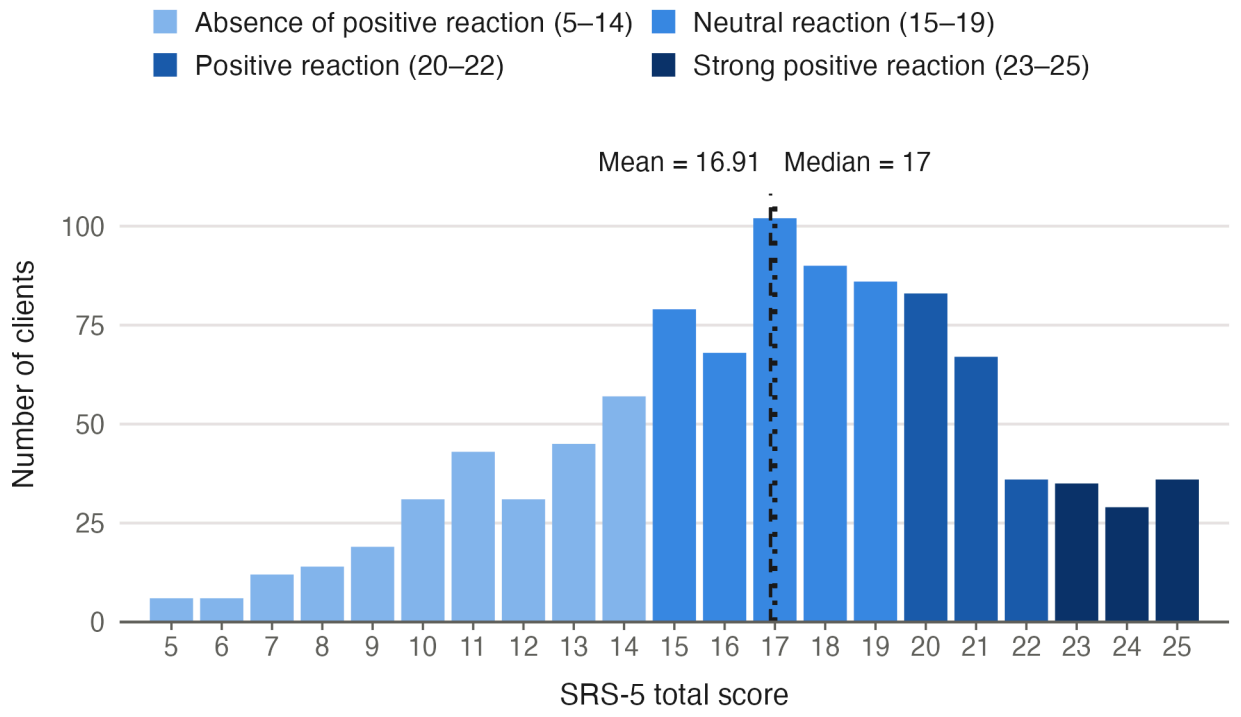
Item	Content	*M*	*SD*	*r*(it)	*r*(ARM)
1	I realised something new about myself or other people	3.38	1.08	.59	.14
4	I feel more empowered, hopeful, or positive about myself	3.36	1.07	.71	.40
7	I have become clearer about the problems or goals for me to work on	3.65	1.03	.66	.38
9	I feel I have improved my skills or learned new strategies to cope with my problems	3.18	1.20	.72	.20
13	I feel emotionally relieved or less burdened	3.34	1.11	.67	.39

Note. M and SD are on the 1–5 response metric. *r*(it) = corrected item–total correlation within the SRS-5; *r*(ARM) = item correlation with ARM-5 total alliance in the matched subsample (n = 68).

Score Distribution and Norms

SRS-5 total scores were approximately normally distributed with a mild negative skew (M = 16.91, SD = 4.39, Mdn = 17, skew = −.32, kurtosis = −.31; Figure 1). Scores spanned the full possible range (5–25). Floor and ceiling effects were minimal (only 0.6% of administrations obtained the lowest possible score and 3.7% the maximum) so the composite avoided the pronounced ceiling effects observed for some individual alliance-type items and functioned as a near-continuous measure. Percentile ranks for each total score are presented in Table 2; for reference, the 25th, 50th, and 75th percentiles corresponded to totals of 14, 17, and 20, respectively.

Figure 1
Distribution of SRS-5 Total Scores (N = 975)



Note. Bars are shaded by interpretive category (see Table 2). Dashed line = mean; dotted line = median.

Table 2
Interpretive Categories and Percentile Ranks for SRS-5 Total Scores (N = 975)

Total score	Item mean & Likert	Percentile rank	Category
5	1 Not At All	0.3	
6	1.2	0.9	
7	1.4	1.8	
8	1.6	3.2	
9	1.8	4.9	
10	2 Slightly	7.4	Absence of Positive Reaction
11	2.2	11.2	
12	2.4	15.0	
13	2.6	18.9	
14	2.8	24.2	
15	3 Somewhat	31.1	
16	3.2	38.7	
17	3.4	47.4	Neutral Reaction
18	3.6	57.2	
19	3.8	66.3	
20	4 Quite a bit	74.9	
21	4.2	82.6	Positive Reaction
22	4.4	87.9	
23	4.6	91.5	
24	4.8	94.8	Strong Positive Reaction
25	5 Very Much	98.2	

Note. Item mean = total ÷ 5. Percentile rank = percentage of the normative sample scoring below each total plus half of those scoring at it (midpoint method). Category cut-points are set at item-mean values of 3.0, 4.0, and 4.5.

Rationale for the Interpretive Categories

Interpretive categories for the SRS-5 were established primarily on the basis of face and content validity rather than purely norm-referenced criteria. Because every item is rated on a verbally anchored five-point scale, the item-mean metric carries intrinsic meaning, and category boundaries were therefore placed at the item-mean values of 3.0, 4.0, and 4.5, which coincide directly with the scale's own Likert descriptors of "Somewhat," "Quite a bit," and half way to "Very much." This yields four bands: an Absence of Positive Reaction band (item mean < 3.0; total 5–14), in which helpful reactions are endorsed less than "somewhat"; a Neutral Reaction band (3.0–3.9; total 15–19) where items are rated as "somewhat" or higher; a Positive Reaction band (4.0–4.4; total 20–22) reflecting reactions that are at least "quite a bit" helpful; and a Strong Positive Reaction band (≥ 4.5; total 23–25) approaching the upper "very much" anchor.

Anchoring the cut-points to the response descriptors keeps the categories transparent and clinically communicable (a clinician can read a client's band directly from the average item response) and avoids defining clinical meaning solely by reference to the normative distribution. Percentile ranks are retained

alongside each score so that this content-referenced banding can be interpreted together with the client's relative standing in the reference sample.

Representation of the Full Scale (Short-Form Fidelity)

Short-form fidelity was examined with a converging set of overlap-aware indices rather than a single coefficient (Table 3). The SRS-5 reproduced the Helpful Reactions construct of the SRS-3-B with very high fidelity. The overlap-free correlation with the five omitted items was r = .81 (66% shared variance), and the correlation with the full 14-item SRS-3-B total was r = .93. Most informatively, the SRS-5 sum correlated r = .98 with factor scores derived from the full 10-item Helpful subscale, recovering 96% of the latent-trait variance, almost as much as the full subscale itself (r = .99). The single common factor accounted for 44.9% of item variance, and the five retained items were, on average, stronger indicators of that factor than the items dropped (mean loading .74 vs. .58), indicating that the short form preserves the best markers of the construct rather than a random subset. To check that this content-based selection did not come at the cost of measurement quality, the SRS-5 was compared with all 252 possible five-item combinations of the ten Helpful items. It ranked 7th of 252 for internal consistency and second for correlation with the latent Helpful factor (r = .98). Internal consistency remained strong (α = .86 vs. .88 for the full subscale; ω = .86 vs. .89), and the Spearman–Brown formula predicted α = .92 had the SRS-5 been lengthened to 10 comparable items, confirming that information was lost only in proportion to length. The short form ordered and classified clients almost identically to the full subscale (percentile-rank r = .96; ICC[A,1] = .96; exact quartile agreement 77.3%, and 99.9% agreement within one quartile). As expected for a measure of helpful impact, the SRS-5 correlated only −.29 with the Hindering Reactions subscale, confirming that it represents the helpful dimension specifically rather than overall session reactivity.

Table 3

Fidelity of the SRS-5 as a Short Form of the Helpful Reactions Subscale (N = 975)

Fidelity index	Value
Internal consistency (Cronbach's α): SRS-5	.86
Internal consistency (Cronbach's α): full Helpful Reactions (10 items)	.88
Internal consistency (McDonald's ω): SRS-5	.86
Internal consistency (McDonald's ω): full Helpful Reactions (10 items)	.89
Spearman–Brown predicted α if SRS-5 lengthened to 10 items	.92
r with full Helpful Reactions total (part–whole, overlap-inflated)	.96 (r^2 = .93)
r with the 5 omitted Helpful items (overlap-free)	.81 (r^2 = .66)
r with full 14-item SRS-3-B total	.93 (r^2 = .86)
Construct fidelity: r of SRS-5 with the latent Helpful factor	.98 (r^2 = .96)
Mean factor loading: SRS-5 items vs. omitted items	.74 vs. .58
Person ordering: percentile-rank r / ICC(A,1) with full Helpful	.96 / .96
Classification: exact quartile agreement / within one quartile	77.3% / 99.9%
Discriminant: r with Hindering Reactions (4 items)	−.29

Note. The part–whole correlation with the full Helpful total is inflated by item overlap and is reported for completeness only; the overlap-free correlation with the omitted items is the unbiased coverage estimate. Construct fidelity = correlation of the SRS-5 sum with factor scores from a single-factor model of all 10 Helpful items. ICC(A,1) = two-way absolute-agreement intraclass correlation between standardised scores.

Convergence With the ARM-5

Of 975 SRS-3-B clients and 10,013 ARM-5 clients, 68 were linked across both measures and contributed a matched pair; 39 of these pairs (57%) were completed on the same day. The SRS-5 correlated r = .41, 95% CI [.18, .59], with ARM-5 total alliance (Spearman ρ = .37; p = .002), a medium effect by Cohen's (1988) benchmarks. Restricting to same-day pairs strengthened the association to r = .63, 95% CI [.39, .79]. At the

component level the SRS-5 related similarly to ARM-5 Bond (r = .37), Partnership (r = .31), and Confidence (r = .31). Of note, the ARM-5 had significant ceiling effects, with 51.5% of administrations in this sample having the maximum score. This ceiling effect limits the correlation strength.

Nonetheless, the magnitude of the correlations between the SRS-5 and ARM-5 indicates substantial shared variance between experienced session impact and the working alliance while leaving the majority of variance unshared, supporting the interpretation that the SRS-5 measures a related but distinct construct.

Discussion

The SRS-5 is a five-item, positively keyed, client-report measure of helpful in-session impact derived from the SRS-3-B. Across a large, naturalistic, predominantly Australian routine-care sample it demonstrated strong internal consistency (α = .86, ω = .86), a well-dispersed and approximately normal score distribution with negligible floor and ceiling effects, and a theoretically coherent, moderate convergent association with an established alliance measure. These properties make the SRS-5 a practical candidate for session-by-session monitoring within feedback-informed treatment, where brevity and repeatability are decisive.

The pattern of convergent validity is informative. The SRS-5 correlated moderately with ARM-5 total alliance, and to a similar degree with each of its Bond, Partnership, and Confidence components. That the measures share about 16% of their variance, while remaining clearly distinct, supports using session impact and alliance as complementary rather than interchangeable indices of session process.

Several limitations temper these conclusions. First, the data are cross-sectional and observational; the present study establishes score structure, distribution, and concurrent convergent validity, but not test–retest reliability, sensitivity to change, factorial structure, or predictive validity for outcome. Second, although large, the sample is a naturalistic convenience cohort drawn from a single platform and from 50 clinicians, who contributed a median of 14 clients each; it is predominantly but not exclusively Australian. Clients of the same clinician may respond more alike than clients of different clinicians, and this clustering was not modelled. Third, both measures showed some negative skew, and the resulting restriction of range is likely to have attenuated the observed correlations, such that true convergence may be modestly stronger. Fourth, the convergent-validity subsample, while based on independent client-level pairs, was small (n = 68) and drawn from clients who happened to complete both measures.

A further limitation concerns what is lost by shortening. The SRS-3 was not developed to measure a single construct; each item represents a type of session reaction that clients themselves regard as important (Řiháček et al., 2024). The SRS-5 takes a different approach, treating the five retained items as indicators of a common helpful-impact dimension, an interpretation supported by its internal consistency and its close correspondence with the latent Helpful factor. The trade-off is that reaction types not retained, including feeling understood by the therapist, acceptance, and the hindering reactions, are no longer assessed separately. The SRS-5 is therefore best used as a brief summary index of helpful impact, ideally alongside an alliance measure, while the full SRS-3-B remains the better choice when item-level feedback on specific reaction types is wanted.

Future work should examine the SRS-5's latent structure and measurement invariance across client subgroups, establish its test–retest stability and responsiveness over a course of therapy, and test whether session-by-session SRS-5 trajectories predict symptomatic outcome and drop-out over and above alliance feedback. Replication in non-Australian and specialist clinical samples would clarify the generality of the present norms. Pending such evidence, the SRS-5 offers a reliable and feasible ultra-brief complement to existing alliance-focused session measures, broadening the process feedback available to clinicians practising feedback-informed treatment.

Concluding Remarks

The case for routine client feedback is now well established: synthesised across trials, giving clinicians structured feedback improves outcomes and reduces deterioration and drop-out (de Jong et al., 2021; Lambert & Shimokawa, 2011). Yet adoption in everyday practice remains stubbornly low. Surveys repeatedly find that only a minority of practitioners administer standardised measures routinely (Hatfield & Ogles, 2004; Ionita & Fitzpatrick, 2014), and even clinicians who endorse the value of monitoring seldom translate that attitude into

practice (Jensen-Doss et al., 2018). The barriers are consistent across studies: the time and administrative burden of measurement, scepticism about clinical utility, and concern that questionnaires intrude upon the therapeutic encounter (Boswell et al., 2015; Solstad et al., 2019). Brevity, on this evidence, is not a cosmetic refinement but a determinant of whether a measure is used at all. Tools that can be completed and reviewed in under a minute are far more likely to be sustained in real-world workflows (Boswell et al., 2015; Miller et al., 2015). By distilling the helpful-reactions construct to five items while retaining 96% of the latent-trait variance of the full subscale, the SRS-5 speaks directly to this barrier: it preserves the information clinicians act upon while minimising the cost that drives their reluctance.

A second, more technical obstacle lies in the measures themselves. Client ratings of the alliance are characteristically high, producing a ceiling effect and a restricted range of scores (Cahill et al., 2012). This is well documented for the Session Rating Scale, whose scores cluster near the maximum (Duncan et al., 2003; Reese et al., 2009), and it was conspicuous for the ARM-5 in the present data, where 51% of administrations reached the highest possible score. Ceiling effects are more than an aesthetic concern: when most clients score at the top, a measure forfeits its capacity to discriminate between sessions, to detect emerging difficulties, and to register change over time; the functions that justify session-by-session feedback. The impact-oriented framing of the SRS-5, which asks what a client took from a session rather than how favourably they regard the therapist, produced an approximately normal distribution with a floor below 1% and a ceiling near 4%. This expanded dispersion should support greater sensitivity to between-session variation and, plausibly, to change, mitigating a constraint that has long limited existing session-feedback instruments.

Taken together, these properties position the SRS-5 as a pragmatic complement to alliance measurement within feedback-informed treatment. It is brief enough to lower the adoption barrier that has hampered routine monitoring, distributed well enough to retain discriminative and potentially change-detecting capacity, and psychometrically faithful to the broader construct from which it was derived. It does not supplant the assessment of the therapeutic alliance (the two are related but distinct), nor the detection of hindering reactions captured by the full SRS-3-B; rather, it offers clinicians an efficient and well-behaved window onto the helpful impact of their work. Whether these advantages ultimately convert into the improved engagement and outcomes that motivate feedback-informed treatment is an empirical question that prospective, longitudinal research should now address.

References

Agnew-Davies, R., Stiles, W. B., Hardy, G. E., Barkham, M., & Shapiro, D. A. (1998). Alliance structure assessed by the Agnew Relationship Measure (ARM). British Journal of Clinical Psychology, 37(2), 155–172. https://doi.org/10.1111/j.2044-8260.1998.tb01291.x

Bordin, E. S. (1979). The generalizability of the psychoanalytic concept of the working alliance. Psychotherapy: Theory, Research & Practice, 16(3), 252–260. https://doi.org/10.1037/h0085885

Boswell, J. F., Kraus, D. R., Miller, S. D., & Lambert, M. J. (2015). Implementing routine outcome monitoring in clinical practice: Benefits, challenges, and solutions. Psychotherapy Research, 25(1), 6–19. https://doi.org/10.1080/10503307.2013.817696

Cahill, J., Stiles, W. B., Barkham, M., Hardy, G. E., Stone, G., Agnew-Davies, R., & Unsworth, G. (2012). Two short forms of the Agnew Relationship Measure: The ARM-5 and ARM-12. Psychotherapy Research, 22(3), 241–255. https://doi.org/10.1080/10503307.2011.643253

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16(3), 297–334. https://doi.org/10.1007/BF02310555

de Jong, K., Conijn, J. M., Gallagher, R. A. V., Reshetnikova, A. S., Heij, M., & Lutz, M. C. (2021). Using progress feedback to improve outcomes and reduce drop-out, treatment duration, and deterioration: A multilevel meta-analysis. Clinical Psychology Review, 85, 102002. https://doi.org/10.1016/j.cpr.2021.102002

Duncan, B. L., Miller, S. D., Sparks, J. A., Claud, D. A., Reynolds, L. R., Brown, J., & Johnson, L. D. (2003). The Session Rating Scale: Preliminary psychometric properties of a “working” alliance measure. Journal of Brief Therapy, 3(1), 3–12.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. British Journal of Psychology, 105(3), 399–412. https://doi.org/10.1111/bjop.12046

Elliott, R. (1985). Helpful and nonhelpful events in brief counseling interviews: An empirical taxonomy. Journal of Counseling Psychology, 32(3), 307–322. https://doi.org/10.1037/0022-0167.32.3.307

Flückiger, C., Del Re, A. C., Wampold, B. E., & Horvath, A. O. (2018). The alliance in adult psychotherapy: A meta-analytic synthesis. Psychotherapy, 55(4), 316–340. https://doi.org/10.1037/pst0000172

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). *irr: Various coefficients of interrater reliability and agreement* (R package version 0.84.1) [Computer software]. https://CRAN.R-project.org/package=irr

Hatfield, D. R., & Ogles, B. M. (2004). The use of outcome measures by psychologists in clinical practice. Professional Psychology: Research and Practice, 35(5), 485–491. https://doi.org/10.1037/0735-7028.35.5.485

Ionita, G., & Fitzpatrick, M. (2014). Bringing science to clinical practice: A Canadian survey of psychological practice and usage of progress monitoring measures. Canadian Psychology, 55(3), 187–196. https://doi.org/10.1037/a0037355

Jensen-Doss, A., Haimes, E. M. B., Smith, A. M., Lyon, A. R., Lewis, C. C., Stanick, C. F., & Hawley, K. M. (2018). Monitoring treatment progress and providing feedback is viewed favorably but rarely used in practice. Administration and Policy in Mental Health and Mental Health Services Research, 45(1), 48–61. https://doi.org/10.1007/s10488-016-0763-0

Lambert, M. J., & Shimokawa, K. (2011). Collecting client feedback. Psychotherapy, 48(1), 72–79. https://doi.org/10.1037/a0022238

McDonald, R. P. (1999). Test theory: A unified treatment. Lawrence Erlbaum Associates.

Miller, S. D., Duncan, B. L., Brown, J., Sorrell, R., & Chalk, M. B. (2006). Using formal client feedback to improve retention and outcome: Making ongoing, real-time assessment feasible. Journal of Brief Therapy, 5(1), 5–22.

Miller, S. D., Hubble, M. A., Chow, D., & Seidel, J. (2015). Beyond measures and monitoring: Realizing the potential of feedback-informed treatment. Psychotherapy, 52(4), 449–457. https://doi.org/10.1037/pst0000031

Norcross, J. C., & Lambert, M. J. (2018). Psychotherapy relationships that work III. Psychotherapy, 55(4), 303–315. https://doi.org/10.1037/pst0000193

R Core Team. (2022). *R: A language and environment for statistical computing* (Version 4.2.0) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Reese, R. J., Norsworthy, L. A., & Rowlands, S. R. (2009). Does a continuous feedback system improve psychotherapy outcome? Psychotherapy*: Theory, Research, Practice, Training*, 46(4), 418–431. https://doi.org/10.1037/a0017901

Revelle, W. (2023). *psych: Procedures for psychological, psychometric, and personality research* (R package version 2.3.9) [Computer software]. Northwestern University. https://CRAN.R-project.org/package=psych

Řiháček, T., Elliott, R., Owen, J., Ladmanová, M., Coleman, J. J., & Bugatti, M. (2024). Session Reactions Scale-3: Initial psychometric evidence. Psychotherapy Research, 34(4), 434–448. https://doi.org/10.1080/10503307.2023.2241983

Solstad, S. M., Castonguay, L. G., & Moltu, C. (2019). Patients' experiences with routine outcome monitoring and clinical feedback systems: A systematic review and synthesis of qualitative empirical literature. Psychotherapy Research, 29(2), 157–170. https://doi.org/10.1080/10503307.2017.1326645

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Appendix

Table A1

Classical Item Properties of the Ten SRS-3-B Helpful Reactions Items (N = 975)

Item	Content	% Not at all	% Very much	Loading	*r*(it)	Decision
1	I realised something new about myself or other people	5.7	14.9	.67	.63	Retained
2	I feel understood, supported, or reassured by my therapist	0.6	65.4	.55	.51	Excluded (A, C)
3	I feel more distanced from certain feelings, thoughts, or memories	21.6	3.8	.38	.36	Excluded (S, L)
4	I feel more empowered, hopeful, or positive about myself	4.6	14.9	.78	.73	Retained
6	I feel more positively or hopeful about another person(s)	15.7	9.0	.64	.60	Excluded (S, O)
7	I have become clearer about the problems or goals for me to work on	2.6	21.7	.73	.68	Retained
9	I feel I have improved my skills or learned new strategies to cope with my problems	9.9	15.8	.78	.73	Retained
11	I feel personally invested in what I need to do in therapy to achieve my goals	1.4	41.7	.55	.51	Excluded (A, C)
13	I feel emotionally relieved or less burdened	5.4	15.8	.75	.70	Retained
14	I have accepted some aspects of myself or my situation more than before	4.9	15.6	.75	.71	Excluded (O)

Note. Items are numbered as in the SRS-3-B. % Not at all and % Very much = percentage of administrations at the lowest and highest response options. Loading = standardised loading on a single factor fitted to the ten Helpful items. *r*(it) = corrected item–total correlation within the ten-item Helpful Reactions subscale. Reasons for exclusion: A = content overlaps with the therapeutic alliance; C = marked ceiling effect; S = reaction relevant only to some sessions; L = weak loading on the common factor; O = content overlaps with retained items.